GB1900: Engaging the public in very large scale gazetteer construction from the Ordnance Survey "County Series" 1:10,560 mapping of Great Britain

Humphrey Southall[1], Paula Aucott[1], Chris Fleet[2], Tom Pert[3], & Michael Stoner[1]

[1] *Department of Geography, University of Portsmouth, Portsmouth, UK*

[2] *Map Library, National Library of Scotland, Edinburgh, UK*

[3] *Data and Technology Team, Royal Commission on the Ancient and Historical Monuments of Wales, Aberystwyth, UK*

Corresponding author: Humphrey Southall – Humphrey.southall@port.ac.uk

Dept. of Geography, Buckingham Building, Lion Terrace, Portsmouth, PO1 3HE, UK

Abstract:

"Semantic" gazetteers should form a geographical "backbone" to our knowledge of the past, but they need to contain toponyms appearing in specific sources, to be under open copyrights and to be large, representing all the features of interest. Academic gazetteer projects are achieving the first two goals but are generally either limited to main settlements or cover small areas. The GB1900 project is creating a new spinal gazetteer of Britain from the Ordnance Survey's County Series 1:10,560 second edition mapping from circa 1900. It will be under Creative Commons licensing. To date, c. 1.7 million features have been identified, and the final result is likely to comprise 3-4 million text strings plus coordinates: possibly the largest specifically historical gazetteer ever built.

This has been achieved through large-scale crowd-sourcing, using Zooniverse-based software. The article describes the project's history, the software system and the transcription process. Over 1,000 volunteer transcribers have been recruited, and the article describes publicity methods, volunteer characteristics and motivation: it is argued that while "citizen science" projects appeal to a general desire to advance knowledge, map-based projects can appeal to more locally-focused individual interests: finding meaning in maps and places. We conclude with recommendations for other similar projects.

GB1900: Engaging the public in very large scale gazetteer construction from the Ordnance Survey "County Series" 1:10,560 mapping of Great Britain

Introduction

This theme issue is focused on semantic historical gazetteers, but "semantic gazetteers" should not be interpreted simply as meaning gazetteers which are represented in the languages of the semantic web, i.e. in RDF or OWL. Defining semantic gazetteers as "URI-based" is substantially more meaningful: in this case, gazetteer entities are identified not primarily by a name or a location but by a Uniform Resource Identifier (URI), meaning a string of text which is not only unique within the particular gazetteer but works as a Uniform Resource Locator (URL) within the World Wide Web (Hart and Dolbear, 2013). This means that identifiers from different gazetteers can be used in combination, or translated between. Many of these potentials are being explored by the Pelagios project (Isaksen Simon, Barker and de Soto Cañamares, 2014).

However, ultimately semantics is not about how information is presented but about what the information means. We also speak of *geo*-semantic approaches in specific contrast to the geospatial, describing the world around us primarily in words rather than through coordinates. This paper is an initial report on the GB1900 project, whose aim is to construct a historical gazetteer of Great Britain in unprecedented detail. This will be made available in a Linked Data format, but that work is yet to be done and will not be of any special technical interest. Rather, its inclusion in this theme issue is justified in three other ways.

Firstly, semantic gazetteers exist primarily not for converting place-names into coordinates but as backbones or spines to which other information is to be linked:

adding meaning to that other information, so we can for example say that a particular "Hereford" in the middle ages had both a fair and a market, whether or not we know exactly where it was; or to draw inferences, such as that "Colwall" was within the ancient Hundred of "Radlow", again without necessarily knowing locations or boundary coordinates. However, before we invest substantial effort in linking our attribute information to such a spine, we need to know that we are not effectively losing control of our data: the gazetteer needs to be in the public domain or under an "open" licence such as Creative Commons. It is now arguably standard practice for historical gazetteers created by academics to be open; as is the widely-used Geonames gazetteer which is available in Linked Data form; as are the large gazetteers created by the U.S government which are the largest single source for Geonames; as is Wikidata, a systematic abstraction from Wikipedia (Vrandečic and Krötzsch, 2014) which provides a spine for the PastPlace linked data historical gazetteer being created by the Great Britain Historical GIS Project (Southall, Stoner and Aucott, forthcoming).

However, none of the above global data sets provide comprehensive and unproblematic listings of every village in Britain, large or small, let alone individual farmsteads or physical features like streams and woods; the same is probably true for most countries. For Britain, several other large data sets exist but none is both open and complete. Firstly, the Ordnance Survey's core MasterMap Topography Layer contains about 470 million individually identified features while their Points of Interest comprise over 4 million *named* features; but these are commercial products. Secondly, their "Open Names" product is freely available under the Open Government Licence, published as Linked Data and contains 2.5 million locations but these comprise "over 870,000 named and numbered roads, nearly 44,000 settlements and over 1.6 million postcodes", so farmsteads, woods and streams are absent. Thirdly, the "Historical

Gazetteer of England's Place-Names" created by the Digital Exposure of English Place-Names (DEEP) project contains four million "historical place-name forms". This is based on research by the Survey of English Place Names, which started in 1923 but is limited to England. Even there only 25 counties are complete, eight are incomplete and seven have yet to be started (Mawer and Stenton, 1924), while the data are not "open", or downloadable. Something new is needed.

Secondly, there is now some agreement that what makes a gazetteer "historical" is not, somehow, the incorporation of a time axis paralleling the spatial coordinates, but the inclusion of names for places which are "attestations": instances of particular text strings being used to refer to the place in particular sources at particular dates; wherever possible, the dated sources should themselves be identified using URIs (Southall, Mostern and Berman, 2011; Grossner, Janowicz and Kessler, 2016). Where the spinal gazetteer is relatively small, as with the Pleiades gazetteer of the ancient world, the extraction of names from specific sources can be done by academic researchers, and that has been the approach of the Pelagios project, noted above. However, if the spine identifies millions of entities, the larger scale approach taken by GB1900 is needed, building links with memory institutions and engaging the general public in crowd-sourcing. To the best of our knowledge, once complete GB1900 will be the largest specifically historical "gazetteer" anywhere.

Thirdly, semantics is about how we find meaning, especially meaning within words. GB1900 is turning a very extensive map into a large corpus of georeferenced text, creating a new and unique object of analysis. The original focus was on place names, but the difficulty of clearly defining resulted in transcribers being asked to extract all text except for some very specific exceptions. These maps have no keys and only limited symbologies, so their meaning lies largely in the text, but conversion to

"georeferenced text" is an essential preliminary to systematic analysis. Ideally the outlines of buildings, fields and so on would also have been captured, but most analyses would probably still simplify them to point coordinates. We believe that the completed open dataset will enable diverse new geosemantic analyses of Britain's historic environments.

The paper begins by surveying previous crowd-sourced gazetteer building projects, and similar. The next section describes the history of the project, and the earlier Cymru1900Wales project on which it builds. The following section describes the maps, the information they contain and, briefly, the project which converted them to an online mosaic. The central section then describes the GB1900 online system, which is not itself a gazetteer as it does not display the transcribed place names, but rather a system for gathering place-name transcriptions. It draws on the infrastructure developed by the Zooniverse team to support very large scale "Citizen Science" projects, notably Galaxy Zoo (Lintott et al, 2008). The final section explores the user experience, covering both the progress of the transcription work and more qualitative material helping us understand why the system, despite some obvious limitations, is successful at its central task. This section is about people finding meaning in maps in another sense: they find the transcription work fun, and repeatedly speak of it being addictive. The short conclusion draws lessons for future projects.

Existing crowd-sourced gazetteer and mapping projects

Digital cultural heritage projects involving online user contributions are now numerous, but we focus here on those working with geographical information, broadly defined.

GeoReferencer, developed by Klokan Technologies, is designed to be easily usable by the general public. It is a web-based application within which volunteers can

geo-reference old maps online by adding points to an on-screen map via comparison with a modern map (Fleet, Kowal and Pridal, P., 2012). Like GB1900, a "leader board" promoted competition between volunteers, and in practice most work was done by a small group of very active contributors.

More closely comparable to GB1900 is the National Library of Wales' Cynefin project (http://cynefin.archiveswales.org.uk/), which asks volunteers to transcribe and geo-reference historical tithe maps for the whole of Wales. Begun in 2014 it has over 1000 registered volunteers and to date has georeferenced 1219 maps and transcribed approaching 1.5 million records. This project is more complicated than GB1900 in that it has four separate transcription elements; transcribing field numbers from the map; transcribing details of the tithe payments from the associated "apportionments"; adding points to geo-reference the maps against out-of-copyright modern and historical base maps and finally clipping completed parish areas to create a single digital layer. However, there is no second transcription required for text elements therefore no inherent verification of accuracy, although the geo-referencing is checked by others adding new or improving existing points.

The AddressingHistory Project run by Edina at the University of Edinburgh georeferenced historical Scottish Post Office Directory records, and crowdsourced their correction and improvement. Parsing and geo-referencing of the ext from the directories was done with the Historical Post Office Directory Parser (POD Parser) software, a bespoke python command line tool created for the project (Osborne, Hamilton, Macdonald, 2014). This code cleans the text strings by using string replaces, stop-words, address lookups and line return fixes and then uses google to fetch coordinates for the addresses. It further assigns an accuracy rating for the address and its API can

create outputs in kml, json & txt format files which incorporate bounding box restrictions.

Lastly, the New York Chronology of Place gazetteer (http://nypl.gazetteer.us/) was created by the New York Public Library in collaboration with Topomancy LLC, partly through crowd-sourcing from historical maps including William Perris's *Maps of the City of New-York* (1852-4) (Vershbow, 2013).

Crowd-sourced transcription and geo-rectification projects such as GB1900 are currently the only feasible way to gather large amounts of textual and spatial place data. However, progress currently being made in automatic vectorisation and Optical Character Recognition (OCR) techniques suggest this is changing. Experiments in Switzerland with a series of algorithms using open source software have produced promising results for some nineteenth century urban maps of Zurich (Iosifescu, Tsorlini, Hurni. 2016). There are limitations to the approach, due to the need to change the specifications for each map series, and sometimes within a series where procedures varied over time.

The Cymru1900Wales and GB1900 projects

What is now GB1900 began as Cymru1900Wales, a project concerned solely with Wales and designed to lay foundations for a survey of Wales modelled on the English Place Names Survey (EPNS) but somewhat faster. Although the main focus of the EPNS is researching the evolution of place names in very early documents, tracing them back to roots in Anglo-Saxon, Norse and Celtic languages, Smith (1954) provides these instructions to EPNS county teams: "The first requirement in the survey of the place-names of a county is a gazetteer", and specifically "a record by parish of all the names on the 6" O.S. map", meaning the Ordnance Survey's 1:10,560 maps described in more

detail in the next section. “This is best achieved by working systematically through each sheet … Slips 8” x 5” used vertically are a convenient size”.

A meeting at the National Library of Wales in May 2011 concluded that the first stage of a Welsh historical gazetteer project should be a similar gathering of names from 1:10,560maps, but that newer technologies should be applied to greatly accelerate the project. In particular, rather than use paper maps, 8” x 5” slips and post-doctoral researchers, crowd-sourcing should be applied to geo-referenced scans of the maps. This suggestion was made by the present lead author and reflected discussions during the multi-day workshop on historical gazetteers held as part of the Association of American Geographers annual meeting in Seattle the previous month; that workshop also led to the publication of Berman, Mostern and Southall (2016) as a major survey of historical gazetteer research.

Fortunately the Royal Commission on the Ancient and Historical Monuments of Wales already had access to geo-referenced scans of the OS 6” maps for Wales, commercially licensed from Landmark Information Group, and licensed for use online in the Welsh Government’s, Peoples’ Collection Wales website. Given the partners’ limited practical experience of crowd-sourcing, the development team behind the Zooniverse crowd-sourcing projects were commissioned to develop software, based on that platform. The funding of this development was secured through contributions from the project partners (National Library of Wales, Centre for Advanced Welsh and Celtic Studies, Royal Commission on the Ancient and Historical Monuments of Wales and People’s Collection Wales). The Cymru1900Wales web site, described in the next section, was launched in October 2013, with publicity via the Welsh Place Names Society and through the National Library’s contacts among local and family history societies. Although the next section discusses the difficulty in obtaining confirmatory

transcriptions, over the first six months over 260,000 transcriptions were added to the system by online volunteers. Activity began to slow mainly because it was becoming harder to find names not already marked.

The most obvious limitation of Cymru1900Wales was its geographical scope, and as the Landmark mapping was a commercial product extension to the rest of Britain would have been very costly. However, in summer 2013 we realized that the progress the National Library of Scotland was making with separate work, described below, to scan and georeferenced the same maps, was creating a way around this limitation. These discussions were initially in connection with new funding bids which proved unsuccessful, but by the summer of 2015 the NLS mapping was complete and the end of Welsh Government funding for the cloud hosting of Cymru1900Wales was nearing, so we decided to move forward without funding.

GB1900 is therefore a joint project of the Great Britain Historical GIS team at the University of Portsmouth, providing developer time and the primary server; the National Library of Scotland, providing access to their six inch mapping; and the Welsh partners contributing the existing software: the Royal Commission on the Ancient and Historical Monuments of Wales, the University of Wales Centre for Advanced Welsh and Celtic Studies, the National Library of Wales and the People's Collection Wales. Each partner, or group of partners, was responsible for publicizing the project in their own national territory. Klokan Technologies Gmbh are, in practice, an additional partner as they provide the cloud hosting for the mapping.

Revisions to the software to turn Cymru1900Wales into GB1900 began in the autumn of 2015. Modifying the system to use the NLS mapping proved trivial but, as discussed below, improving the transcription process and moving the system out of the cloud took longer. An essentially complete version of the software was presented at the

Digital Humanities conference in July 2016. On 6th September 2016, Cymru1900wales was closed down, its database of volunteers and transcriptions moved transferred to the GB1900 system and that system went live. However, it was initially publicised only to the existing Welsh volunteers. Interestingly, they discovered that the higher-resolution NLS mapping revealed substantial numbers of names and other text strings within the Welsh part of the mapping still needing to be transcribed.

The public launch was on September 22nd. Publicity included a press release to English regional media but this had limited impact: we know only of one newspaper article in Cornwall and the lead author being interviewed on BBC Radio Sheffield; and there was as almost complete absence of subsequent transcription activity in the Sheffield area. Other publicity was via social media, specifically Facebook and Twitter; via internet mailing lists targeted at local historians, archivists and map librarians; and, with some time delay, specialist newsletters. We were unable to obtain funding for printed publicity materials, which we would have distributed primarily to local history libraries and archives.

Ordnance Survey County Series mapping

The Ordnance Survey (OS) six-inch to the mile (1:10,560) County Series is the most detailed topographic mapping that covers all of Scotland, England and Wales from the 1840s to the 1950s. The OS published more detailed scales of mapping, at 25 inch to the mile (1:2,500) covering most settled rural areas, and at five or ten feet to the mile (1:1,056/1:500) for settlements with more than 4,000 people. However, the six-inch was the most detailed scale that covered all areas. Following the original survey between 1842-1893, the mapping was revised for the whole country between 1888-1914, and then updated regularly between 1914 to the 1940s but only for urban or rapidly

changing areas. There are therefore just two editions providing comprehensive coverage. Like Cymru1900Wales, GB1900 uses the second edition. One advantage is that the date range of sheets across the country is shorter. Another is that the creation of a seamless layer is easier, as the first edition maps have map coverage just up to county boundaries, the rest of the sheet being blank, so creating a seamless layer with first edition maps is more labour intensive.

The second edition maps are easier to seam together for geodetic reasons. The first edition used a different projection origin for each county, so even if the blank areas were not a problem sheets would not fit together at county boundaries, without extensive warping or rubber sheeting. From the 1870s, OS successfully reduced this problem by combining counties together on the same meridian. 14 counties in central England shared the Dunnose origin on the Isle of Wight, whilst 11 counties in or bordering on Wales shared the Llangeinor origin in South Wales, and when Scottish county maps were revised in the 1890s, many were combined onto the same origins too (Adams, 1989-90).

The OS six-inch maps are an excellent record of practically all man-made and natural features in the landscape. Although many of these features were included on earlier estate, county, tithe and military maps, many landscape features and place names make their debut on these maps due to the comprehensive remit of the OS, and the large manpower employed in the survey work. The maps show details of communications such as roads, railways, lanes, tracks and canals, rural landscape features such as fences, walls, fields, streams, and farms, and the characteristics of settlements, including the names of many public buildings, such as inns, hotels, public houses and industrial premises. Although urban detail is generalised, even relatively small features, such as letter boxes, bollards on quaysides, mile posts, and flag-staffs are all shown, as is the

permanent detail of quarries, pits, slag heaps and refuse tips, so forming a valuable record of environmental risk and potential contamination today. Uncultivated land is distinguished by over ten different symbols, including those for different types of woodland (e.g., birch, fir, mixed woodland, furze, osiers, and brushwood), as well as marsh, bog, and rough grassland. The OS had a clear remit to record all administrative boundaries, including civil parish, burgh, and county boundaries, but this was not completed for the first edition. The second edition six-inch maps are therefore the most detailed countrywide record of administrative boundaries following the extensive revisions of the Local Government Acts of 1889-1894. The surveying work also involved detailed levelling, recording spot heights in feet above mean sea level at Liverpool, and the six-inch maps are the most detailed scale at which contours are shown (Oliver, 2013).

Of course, the place names recorded on the six-inch maps were and are today one of the most useful and important elements of them for many purposes, and OS had a very clear and formalised procedure for determining what they felt should be recorded as a single "correct orthography" of names.

> "For the name of a house, farm, park or wood, or other part of an estate the owner is the best authority. For names generally the following are the best individual authorities and should be taken in the order given: Owners of property; estate agents; clergymen, postmasters and schoolmasters, if they have been some time in the district; rate collectors; borough and county surveyors; gentlemen residing in the district; Local Government Board Orders; local histories; good directories. Assistance may also be obtained from local antiquarian and other societies, in connection with places of antiquarian and national interest. Respectable inhabitants of some position should be consulted. Small farmers and cottagers are not to be depended on, even for the names of the places they occupy, especially as to the spelling. But a well-educated and independent occupier is, of course, a good authority" (OS Instructions to Field Examiners (1905) quoted in Seymour, 1980, 176).

The place names on the maps, as for other features, therefore reflect very deliberate policies for preferring particular forms over others, and often erasing local vernacular forms, as well as generally preferring English forms over Welsh and Gaelic (Harley & Walters, 1982; Withers, 2000). The names also directly depended upon what were considered the "best authorities" within reach of the mapping party and, in turn, upon the authorisation of these authorities' views in written form in the OS name books.

The National Library of Scotland originally collected only map sheets covering Scotland. However, in the late 1960s, when the OS needed to pass on their record collection of maps at Burley in Hampshire, NLS was the main recipient, and these holdings were supplemented by further donations in 2009 with the move of OS from Romsey Road. It is estimated that of the ca. 47,000 OS six-inch maps covering Scotland, England and Wales, NLS lacks around 37 sheets or 0.1 % and so the NLS holdings are the most comprehensive set held outside the British Library and Bodleian Library.

Due to resource limitations, work to put the maps online was spread over at least a decade; first edition six-inch mapping of Scotland went online in 2004, second and later edition mapping in 2009, and all editions of England and Wales in 2014. The main workflow involved listing the sheets and their dates prior to scanning, then georeferencing the maps, putting online both individual sheets as zoomable images and a georeferenced mosaic of each edition (Fleet & Pridal, 2012; Fleet, 2014). Geographic metadata, in the form of shapefiles of County Series sheet boundaries and numbers, was obtained through Ed Fielden's Coordinate Convertor. Fortunately, apart from some first edition mapping, the NLS sheets are not bound, and therefore sheet-feed scanners were used to capture the images at 400 dpi in 24 bit colour (three 8-bit channels of red, green,

and blue). This quality is demonstrably superior to the 300 dpi bi-tonal scans created by the Landmark Information Group in the 1990s, as formerly used by Cymru1900Wales, and this quality differential was one of the main rationales for replicating the scanning work. Cropping and geo-referencing of each sheet used the MapSheetAutoGeoRef plugin in QGIS, linking the four corners of each map to its shapefile polygon coordinates. The six-inch 1888-1914 layer is made up of 19,165 sheets, and Klokan Technologies' MapTiler software was used to prepare the resulting GeoTIFFs as a single tileset for rapid online delivery. Klokan Technologies' Tileserver CDN was used to host the six-inch layer used by GB1900, providing robust, scalable and fast access to the layer.

It should also be noted that presenting these maps as a seamless georeferenced layer creates some anomalies, partly due to names for features covering a wider geographic area (such as jurisdictions, districts rivers, etc.) that are replicated across sheets, and partly due too, to the overlaps between sheets at county boundaries. Whilst the OS generally planned the presentation of names across sheets consistently across a whole county as a set to reduce replication and allow some names to run across sheets, where sheets from different counties meet this is often not the case, and replication (or sometimes truncation) of names can occur. A related problem can occur where sheets from adjacent counties overlap. The original layer used by Cymru1900Wales from Landmark Information Group sometimes used different sheets at county boundaries from those used by NLS, resulting in transcribed features that had no cartographic basis on some county boundaries in Wales. It has been possible to address this problem by presenting the specific sheets used by Cymru1900Wales as a special overlay layer in GB1900 as a partial fix. More generally, in addition to the selection process for the

names themselves, it is another reason why the project team are careful not to present GB1900 as a "definitive" set of 1900 place names.

The Cymru1900Wales and GB1900 systems

Like most modern web applications, the Cymru1900Wales application server was built from many existing components, both software libraries and cloud-hosted services. It was hosted on Heroku, a cloud Platform-as-a-Service supporting several programming languages designed for web application deployment. Its operating system was Ubuntu Linux, running within a virtual machine (VM). Web applications to be deployed on Heroku are expected to be held in a separate Git-based repository, then "buildpacks" supplied by Heroku transform the repository's application code into an executable package built specifically for a particular VM image, or "stack".

The Cymru1900 application was written in Ruby on Rails, a server-side web application framework written in the Ruby language. "Rails" is a model–view–controller (MVC) framework, providing default structures for a database, a web service, and web pages. The database used was MongoDB, a "NoSQL" JSON store database, which means that on the one hand it is particularly easy to hold components for a web page, but on the other that the team cannot use the familiar SQL language for manipulating the data, or the spatial functionality the major relational databases provide. The Rails Web App uses Mongo Mapper, an Object Relational Mapper (ORM) for Ruby, to interact with MongoDB. Volunteer registration, including for example the management of lost passwords, used "devise", a flexible authentication solution for Rails based on the Warden authentication framework. Devise uses another software component, mm-devise, to interactive with MongoMapper. The map tiles from Landmark Information were hosted on Amazon S3 (Simple Storage Service).

GB1900 was always conceived as an extension of Cymru1900, so software development addressed a series of specific issues. Firstly, modifying the system to work with the National Library of Scotland's web map server proved trivial. Klokan Technologies' Tileserver platform makes the NLS OS six-inch layer available as an Open Geospatial Consortium Web Map Tile Service. The OS six-inch layer could therefore easily be used inside the GB1900 Leaflet-based web-mapping system. The system could equally easily use other georeferenced map layers that are made available using similar open OGC standards, a fact which could well be useful for crowdsourcing names from other historic maps in future.

Secondly, GB1900 has no actual funding so the monthly cost of Heroku hosting was problematic. The GB Historical GIS team had a new server available which is now also running the Vision of Britain through Time web site but which was available for the sole use of GB1900 during the initial launch. Given that Git remained the actual code repository, redeploying the application onto this server was in itself a simple procedure, but required the team to address software dependencies which were becoming increasingly problematic regardless of where the system was hosted. The largest was that MongoMapper had not been updated since early 2012 and was incompatible with current versions of Ruby on Rails. Another was that changes to the leader board, discussed below, required new functionality available only in newer versions of MongoDB, raising further compatibility issues. That said, hosting on our own server rather than in a cloud-hosted VM makes it somewhat easier to keep running old versions of components.

[Figure 1 near here]

Thirdly, the largest changes made were necessarily to the user interface. As discussed above, the Cymru1900Wales system was very effective in gathering initial

transcriptions, but few were ever confirmed. When a volunteer transcribes text on the map, an orange pin icon would represent this location and their transcription. When a logged-out volunteer went to the site the pins were grey, with no information on how many transcriptions had been made. As the screenshot in figure 1 shows, when a volunteer logs in the grey pins transform into a grey pin with an attached bubble showing whether "1" or "2" additional transcriptions were required before it was marked complete. When a volunteer adds a transcription to a grey pin it would be transformed into an orange pin, without the attached additional transcriptions bubble. The orange pin solely identifies to that volunteer they had contributed to it. If the transcription of the text was confirmed, and the current volunteer had not contributed the pin would be red. This meant that volunteers and potential new volunteers could not scan the map to quickly identify the overall progress of the project.

A key decision for GB1900 was to reduce the requirement for confirmations from three independent transcriptions to also allow confirmation via only two matching transcriptions. We also added a little flexibility to the string matching; for example, as volunteers transcribe the text 'F.P.' for footpath variously as 'FP', 'F.P', or 'F.P.', these full stops would be ignored.

[Figure 2 near here]

As figure 2 shows, GB1900 makes greater use of different colored pins, and displaying these to potential volunteers which meant three colors was sufficient to provide complete information without the pop-ups. It was originally felt that a "traffic light" set of red, amber and green would best communicate the state of a transcription, but these colors are not color blind-safe so instead the system uses green, brown and purple. The specific colors were chosen using http://www.somersault1824.com/tips-for-designing-scientific-figures-for-color-blind-readers.

- Green pins mean further work is needed, either because they have been transcribed only once by another volunteer or because the second transcriptions did not match the first, thus requiring a third.
- Brown pins are ones the logged-in volunteer can do no further work on, either because they were the initial transcriber or because they have done a "confirmatory" transcription which did not match.
- Purple pins are complete, so nobody can do further work on them.

The final result is that there are two distinct work flows for volunteers. When making initial transcriptions, one looks for text strings on the map which are not already marked by any pin: clicking next to them brings up a pop-up dialogue into which the string can be typed, and clicking on done or pressing return creates a new brown pin. When confirming, one looks for green pins and clicks on them to bring up the dialogue box: typing in the name given on the map where the pin is located then clicking on done. Confirmations will always change the pin's color, to purple if one's transcription matches an existing one or to brown otherwise. Note that the revised interface does not tell the volunteer how many attempts have been made, only whether or not two of them have matched. When another volunteer correctly transcribes the text relating to this pin, it will change to be a purple pin for all volunteers. So the aim of the project is to have a purple pin for each piece of text on the map.

Fourthly the "leader board" which identifies those who have made the largest contributions by displaying the "top ten" has been revised. In Cymru1900 ranks were based entirely on numbers of initial transcriptions, but in GB1900 it is based on whichever is smaller of the number of initial and of confirmatory transcriptions; as noted above, this comparison required a later version of MongoDB. This approach seems to have worked as there is now an active group of contributors who are clearly

alternating the two tasks. However, at the time of writing it would take nearly 50,000 transcriptions before a new contributor would appear on the leader board, and it would be helpful if there were also leader boards for particular parts of Great Britain. We have not found a computationally efficient way to tell contributors their personal rank, if they are not in the top ten.

Engaging users

The success of GB1900 clearly requires large-scale user engagement, and although the two national libraries to some extent had existing publicity channels reaching relevant groups in Scotland and Wales, the GB Historical GIS team had no English equivalents. A press release was sent out to English regional newspapers and broadcasters, but this seems to have led only to one article in Cornwall and a local radio feature in Sheffield; and there was no evidence whatsoever of any consequent increased transcription activity in these areas. Announcements were sent to various internet mailing lists, notably those for archivists, map librarians and local historians, which did lead to requests for newsletter articles going out to the British Association for Local History and to the Society for One-Place Studies. Lastly, tweets were put out by the two libraries and via @gbhgis; but the latter was almost moribund with few followers. Nothing specific was done on Facebook, and despite it being Zooniverse-based, GB1900 like Cymru1900 was not a Citizen Science Alliance project so had no access to their existing network of volunteers.

[Figure 3 near here]

However, GB1900 has been very successful in engaging users. Figure 3 shows numbers of transcriptions, confirmations and new user registrations per week. In the first month

following the launch 325 new volunteers registered with the site; on average, over the first four months since launch, each day 39 volunteers have been active, creating 9,807 transcriptions and 7,173 confirmations. At the time of writing, there are over 2.5 million total transcriptions, of which the top ten individual transcribers are responsible for a million.

This success follows not from our limited publicity work but from the fascination maps hold for people, especially when they are maps of places which have meaning for them, coupled with the software being genuinely easy to use without training. The clearest evidence for this comes from tweets which either contain #gb1900 or gb1900.org.

Firstly, many comments describe the experience as pleasurable or, especially, addictive: "You know the thing you should be doing? Well you won't being doing it coz you're now addicted to http://GB1900.org" (26 Sept); "have you become a #volunteer gb1900 transcriber yet? the next addiction :)" (8 Oct); "Strangely addictive helping to crowdsource place names on 1900 O.S. map: http://www.gb1900.org/ Try it for somewhere you know!" (5 Nov); "Trying to fit in some typing on http://GB1900.org before the school run" (17 Jan); "Have done much of my home town, and that of my ancestors. Found it really addictive" (17 Jan).

Secondly, and almost certainly in contrast to most volunteers working on citizen science projects such as Galaxy Zoo, our volunteers were motivated by the particular places they worked on: "I found out about GB100.org from Twitter, and went across to have a look. Well, more than just a look: I ended up spending the evening adding entries for my locality. It gets to be surprisingly addictive when the places you are logging are ones that you know." (27 Sept; blog post at http://light.demon.co.uk/wordpress/the-gb1900-org-project-first-look/); "So far on #gb1900 I've found Hades Hill, Hunger Hill

Lane, Hailstorm Hill & Mucky Earth near my #oneplacestudies. Maybe life was hard?" (18 Oct); "I've been working my way up Leicestershire this evening" (17 Jan); "Transcribing the 1901 census addresses in my family tree. Slow but fun!" (17 Jan).

That geography matters is also visible in individual transcription strategies visible on the site itself. Some people obviously had an interest in a particular feature in the landscape, for example all the mills in Burnley were labelled within the first few days, but it took a couple more weeks for anything else to be transcribed within the town. Other volunteers appear to follow a route across the countryside, often along footpaths, although at least one person concentrated on the footpaths which followed routes now used by modern motorways.

Thirdly, the software matters, and the improvements to the transcription process itself probably mattered more than the leader board: "Very easy to use and will create valuable reusable data. Join us!" (4 Oct); "Ok I admit it, sometimes I AM motivated by a leader board" (5 Oct); "the website works well on tablets! Perfect for bedtime" (17 Jan); "So satisfying when you find an area no-one else has touched and all the pins are brown" (17 Jan).

[Figure 4 near here]

During the registration process volunteers are asked three questions about their background to assist with user analysis, covering gender, age range and how they heard about the project. Each question gives users the option not to answer. These registration process questions do not include about 120 volunteers who signed up via their Facebook log-in, as the metrics for these registrations are collected differently. We were obviously concerned not to overburden new recruits, so we are planning a more detailed online questionnaire which we invite volunteers to complete.

Figure 4 shows answers to the questions on age and on how they discovered the site. During the first month there was a fairly even spread of recruits within the two more senior age bands attracting the most volunteers, 50-64 (74) and 65 and over (78). The following two months saw a greater number joining the over 65's category, but the most recent month has seen more volunteers from the middle age bands of 50-64 (27) and 25-49 (18). In terms of methods of discovery there was a fairly even split between mailing lists, blog or forum posts and other online mediums during the first month. Facebook has less, but as volunteers can register directly via a Facebook App this statistic is a little misleading. Since then the discovery methods seem to predominately be other online and offline sources, including word of mouth and other written sources which may include the newsletter articles. In hindsight a reference to Twitter as a means of discovery would have been a useful addition to the methods suggested.

[Figure 5 near here]

Figure 5 shows the geographical progress of the transcription process. The GB1900 transcription tool can only show transcription markers when quite tightly zoomed in, and even then becomes very slow in areas of dense transcriptions. These maps were created in a separate online system at http://geo.nls.uk/maps/gb1900, which uses GeoServer and GeoWebCache to create pre-rendered maps at different scales, enabling the project team and contributors to identify areas needing more work. All maps show initial transcriptions, not confirmations.

At launch, Wales was almost complete due to data inherited from Cymru1900, while England and Scotland were blank. The first map shows coverage after a month, with almost complete coverage of Wales plus dense patches elsewhere, such as Oxfordshire. Most of the map is empty, but it is very visible that some users were following transport routes. After two-and-a-half months, large areas of southern and

eastern England and central Scotland were densely transcribed, but progress in northern England and the south-west has been slower. Even in the final map, showing progress at the time of writing, after four months, the north-south divide remains and the effects of individual transcription strategies are still visible.

In conclusion, it seems that if you build a sufficiently compelling map-based crowd-sourcing system, the volunteer transcribers will come, somehow arriving despite limited publicity. One way this has happened is that three months into the project we were approached by *Who do you think you are?*, a popular magazine for family historians linked to the television program of the same name. They were organising their first Transcription Tuesday as an event for January 17, encouraging their readers to contribute to one of six online projects. Each was advocated by a member of the editorial team, and we were the choice of the editor herself. We were the second most successful project in terms of new transcribers recruited, and it led to our single most successful day, with 25,316 transcriptions plus 18,493 confirmations. At the time of writing we are seeing a second mid-project boost through publication of the British Local History Association article.

Conclusion

[Figure 6 near here]

This article is about the project, and the process of user-engagement, rather than about the results: the project remains very much in progress, with coverage of England and Scotland far from complete. However, Figure 6 is essentially a result from the earlier Welsh work and shows how the project is starting to fulfil the original goal of creating a welsh equivalent to the Survey of English Place Names. In Welsh, both

“Hafod” and “lluest” refer to summer pastures whilst “pentre(f)” in modern standard Welsh generally means a village, although in some areas it is known to refer to substantial farms and perhaps this meaning was more prevalent across Wales at an earlier date. Figure 6 shows that “pentre” appears mainly in lowland areas while both “Hafod” and “lluest” appear mainly in uplands, but the particular concentration of “lluest” in central Wales requires further investigation.

Although that example concerns repetitive elements within essentially unique toponyms, longer-term much of the meaning in the maps lies in repeated strings identifying types of feature, for example environmental hazards such as land marked as “Liable to floods”, or the distribution of cultural markers, such as the churches and chapels of different denominations. The output files mapped in Figure 5 are, very clearly, instances of geographical “Big Data” and we expect a range of new analytical uses to emerge as they go into general circulation, even as we create cleaned gazetteers limited to more narrowly defined place names. However, the remainder of these conclusions provides recommendations to other projects seeking to build large gazetteers through crowd-sourcing.

Firstly, the overall history of both Cymru1900 and GB1900 shows that it is enormously important for any online project requiring user interaction to have some capacity to revise the software post-launch: the original software was very successful at getting initial transcriptions, especially in the six months following launch, but poor at getting confirmations. Even though Cymru1900 was kept running for just under three years, right up to the GB1900 launch, few Welsh names were ever confirmed, but the revised transcriptions interface and “leader board” have solved this problem.

A related problem that we have not been able to address is that Cymru1900/GB1900 is a web app, meaning a Javascript program which runs in a web

browser, but not a web site: everything which appears on-screen is created by the program. The practical consequence is adding any information at all requires work by a software developer so other members of the project cannot provide advice or feedback to contributors. We have now added both a separate "support site" and the visualisation system behind figure 5, but our capacity to engage with users remains limited, and mainly via Twitter.

Thirdly, technical problems with the map sheets themselves have posed significant practical problems for users: differences between the map sheets used in Cymru1900Wales and in GB1900, especially along county boundaries; a very few sheets out of 19,165 being simply missing. Here again it was very important that for some months following launch we had a developer available to revise the system.

Despite these problems, the system has clearly been very successful in its central task, in two senses. Firstly, and as discussed above, people enjoy doing this, and much of this is because it is a way of closely engaging with old maps and with places of significance to them. This is arguably a significantly different motivation from mainstream Citizen Science: most people do not have personal favourite galaxies (Raddick et al, 2008). Secondly, the infrastructure has proved capable of handling large numbers of concurrent users. Particularly notable is how the pins being added by one user will appear in the browser of another user working on the same area, without that user taking any action. We hope that other academic projects will start to similarly engage with the general public on a large scale.

Finally, although this paper mainly concerns creating just a specific component of semantic gazetteers, there is perhaps a broader conclusion concerning the broader area of geo-semantics. We normally think of maps as being collections of points, lines and polygons, but they also contain a very large quantity of text, and it is the text which

provides most of the meaning in a map; this is particularly true of the County Series, where the individual sheets contained no key. In GB1900, we are extracting no lines or polygons, only points, but we are also extracting almost the entire semantic content of the maps, and so creating a new and primarily textual representation of Britain's geography a century ago.

List of websites

Co-ordinate Converter: http://www.fieldenmaps.info/cconv/

Cymru1900Wales: http://www.cymru1900wales.org/

GB1900: ***http://www.gb1900.org/***

GB1900 support site: https://support.gb1900.org/

GB1900 output visualisation: http://geo.nls.uk/maps/gb1900/

Leaflet tileLayer documentation: http://leafletjs.com/reference.html#tilelayer

MapTiler: https://www.maptiler.com/

NLS Historic Maps Subscription API: http://maps.nls.uk/projects/subscription-api/

NLS Map Images website: http://maps.nls.uk

Tileserver: https://tileserver.com/

OGC Web Map Tile Service standard: http://www.opengeospatial.org/standards/wmts

References

Adams, Brian. 1989-1990. 198 years and 153 meridians, 152 defunct. Sheetlines 25-27. https://www.charlesclosesociety.org/files/153Meridians.pdf

Berman, M., Mostern, R. and Southall, H.R. (eds.) 2016. Placing names: enriching and integrating gazetteers. Indianapolis: Indiana University Press.

Chiang, Y.; and Knoblock, C. A. 2015. Recognizing Text in Raster Maps. GeoInformatica, 19 1-27. http://link.springer.com/article/10.1007%2Fs10707-014-0203-9

Fleet, C., Kowal, K.C. and Pridal, P. 2012. Georeferencer: Crowdsourced georeferencing for map library collections. D-Lib Magazine, 18(11): 5. http://www.dlib.org/dlib/november12/fleet/11fleet.html

Fleet, C., Pridal, P. 2012. Open source technologies for delivering historical maps online - case studies at the National Library of Scotland. LIBER Quarterly 22: 240-257. https://www.liberquarterly.eu/articles/abstract/10.18352/lq.8052/

Fleet, C. 2014. Old maps and new web technologies: practicalities, impact and potential. Society of Cartographers Bulletin 48: 26-34.

Grossner, K., Janowicz, K. and Kessler, C. 2016. Place, period and setting for Linked Data gazetteers. In Berman, M., Mostern, R. and Southall, H.R., eds. Placing names: enriching and integrating gazetteers. Indianapolis: Indiana University Press, 80-96.

Harley, J. B., Walters, Gwyn. 1982. Orthography and Ordnance Survey Mapping 1820-1905. Archaeologia Cambrensis 131: 98-135.

Hart, G., and Dolbear, C. 2013. Linked data: a geographic perspective. Boca Raton: CRC Press.

Iosifescu, I., Tsorlini, A., Hurni, L. 2016. Towards a comprehensive methodology for automatic vectorization of raster historical maps. e-Perimetron, 11(2): 57-76 http://www.e-perimetron.org/Vol_11_2/Iosifescu_et_al.pdf

Isaksen, L., Simon, R., Barker, E.T.E., and de Soto Cañamares, P. 2014. Pelagios and the emerging graph of ancient world data. In Proceedings of the 2014 ACM conference on Web science (WebSci '14). ACM, New York, NY, USA, 197-201.

Lees, L.H. 1979.Exiles of Erin: Irish migrants in Victorian London. Manchester: Manchester U.P.

Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Szalay, A., Andreescu, D. and Murray, P., 2008. Galaxy Zoo:

morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society, 389(3):1179-1189.

Mawer, A. and Stenton, F.M. eds., 1924. Introduction to the Survey of English Place-names. Cambridge: The University Press.

Oliver, R. 2013. Ordnance Survey maps: a concise guide for historians, Third ed. London: Charles Close Society.

Osborne, N., Hamilton, G. & Macdonald, S. 2014. Historical Post Office Directory Parser (POD Parser) Software from the AddressingHistory Project. Journal of Open Research Software. 2(1): 23. DOI: http://doi.org/10.5334/jors.aq

Raddick, M.J., Bracey, G., Gay, P.L., Lintott, C.J., Murray, P., Schawinski, K., Szalay, A.S. and Vandenberg, J., 2009. Galaxy zoo: Exploring the motivations of citizen science volunteers. arXiv preprint arXiv:0909.2925.

Seymour, W.A. (ed.) 1980. A history of the Ordnance Survey. Folkestone, Kent: Dawson.

Smith, A.H. 1954. The Preparation of County Place-Name Surveys. London: English Place-Name Society.

Southall, H., Mostern, R. and Berman, M.L. 2011. On historical gazetteers. International Journal of Humanities and Arts Computing 5(2): 127-145.

Southall, H.R., Stoner, M. and Aucott, P. forthcoming. PastPlace historical gazetteer. Article in Huang, Bo (ed.) Comprehensive Geographic Information Systems. Oxford:Elsevier.

Vershbow, B., 2013. NYPL Labs: Hacking the library. Journal of Library Administration, 53(1): 79-96.

Vrandečić, D., and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. Communications of the ACM CACM, 57(10): 75-85.

Withers, C.W.J. 2000. Authorizing landscape: ‘authority', naming and the Ordnance Survey's mapping of the Scottish Highlands in the nineteenth century, Journal of Historical Geography 26: 532-54.

List of Figures

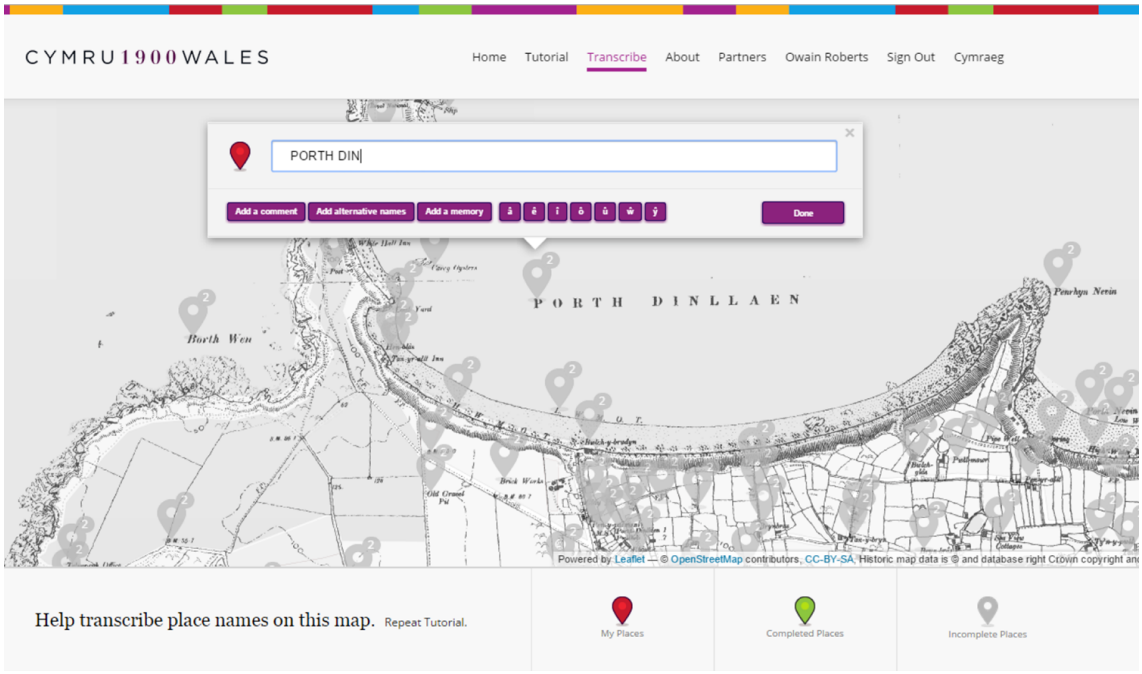

Figure 1. Place-name transcription interface in Cymru1900Wales.

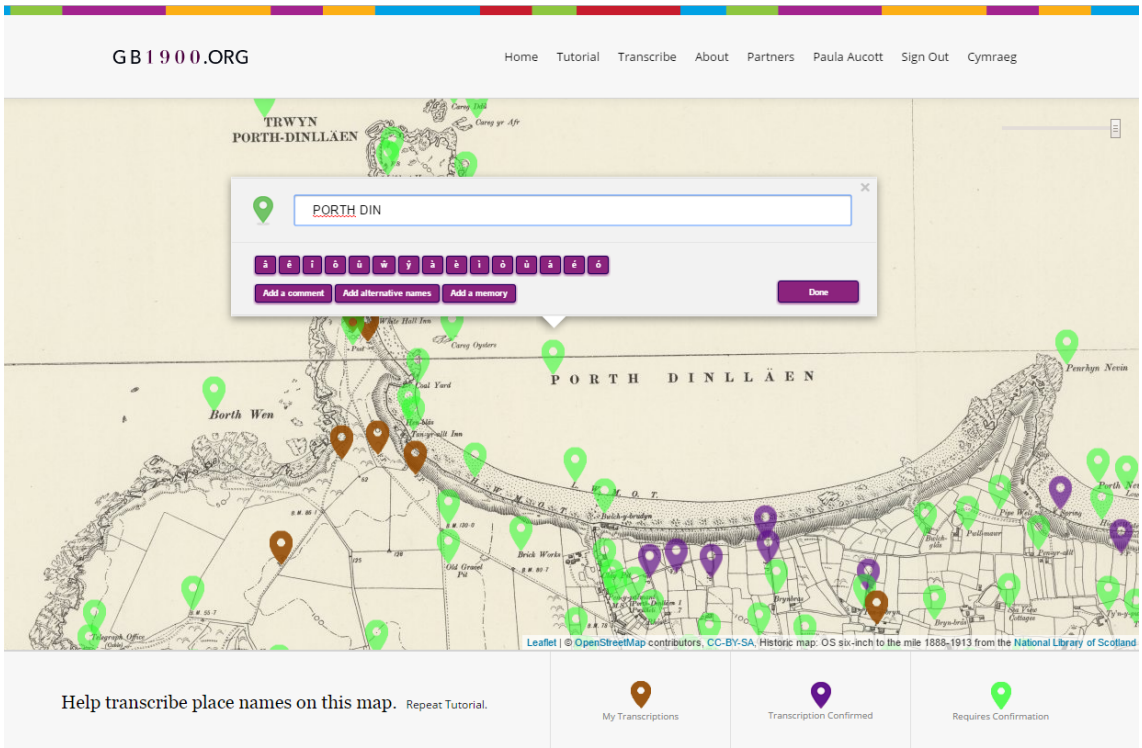

Figure 2. Modified place-name transcription interface in GB1900.

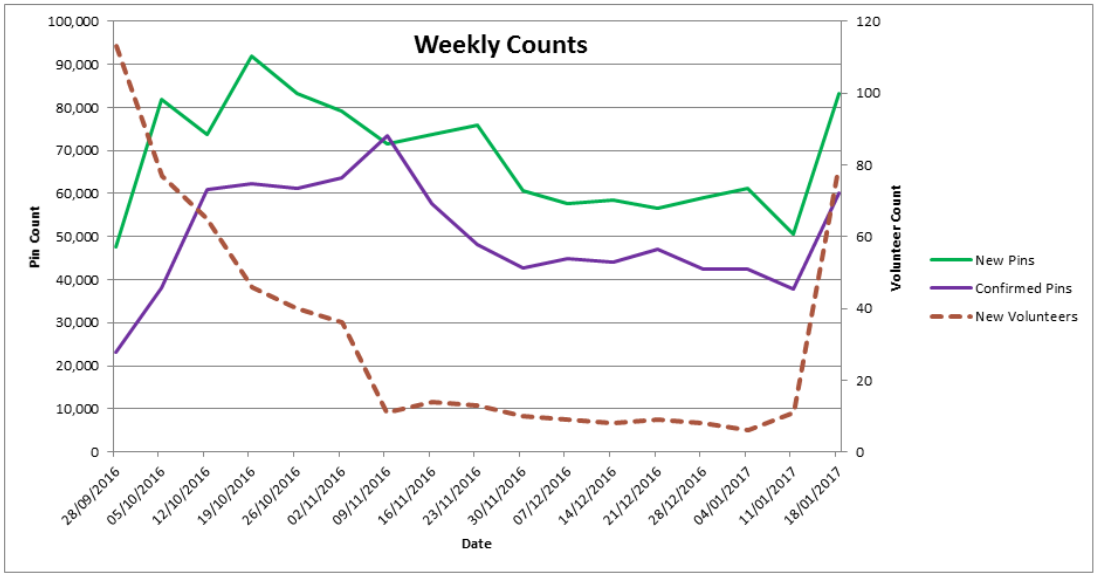

Figure 3. Numbers of GB1900 transcriptions, confirmations and volunteers over time.

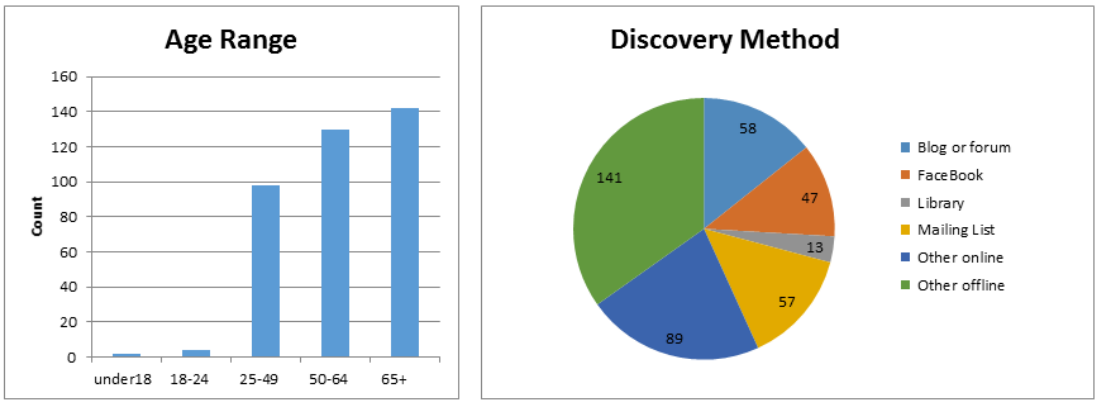

Figure 4. Characteristics of GB1900 volunteers.

Figure 5. Maps showing progress of transcription: (a) 20th October 2016.

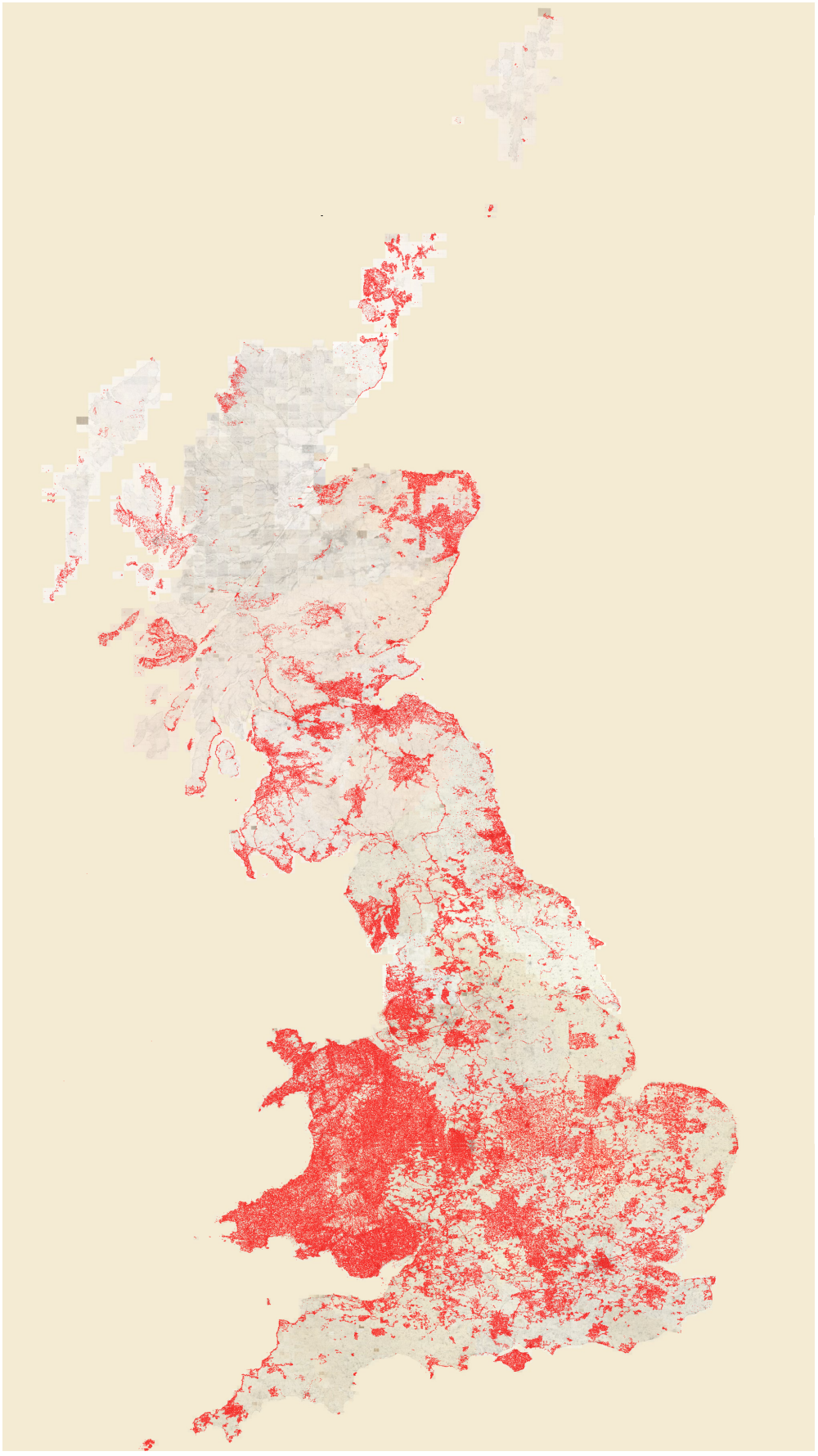

Figure 5. Maps showing progress of transcription: (b) 2nd December 2016.

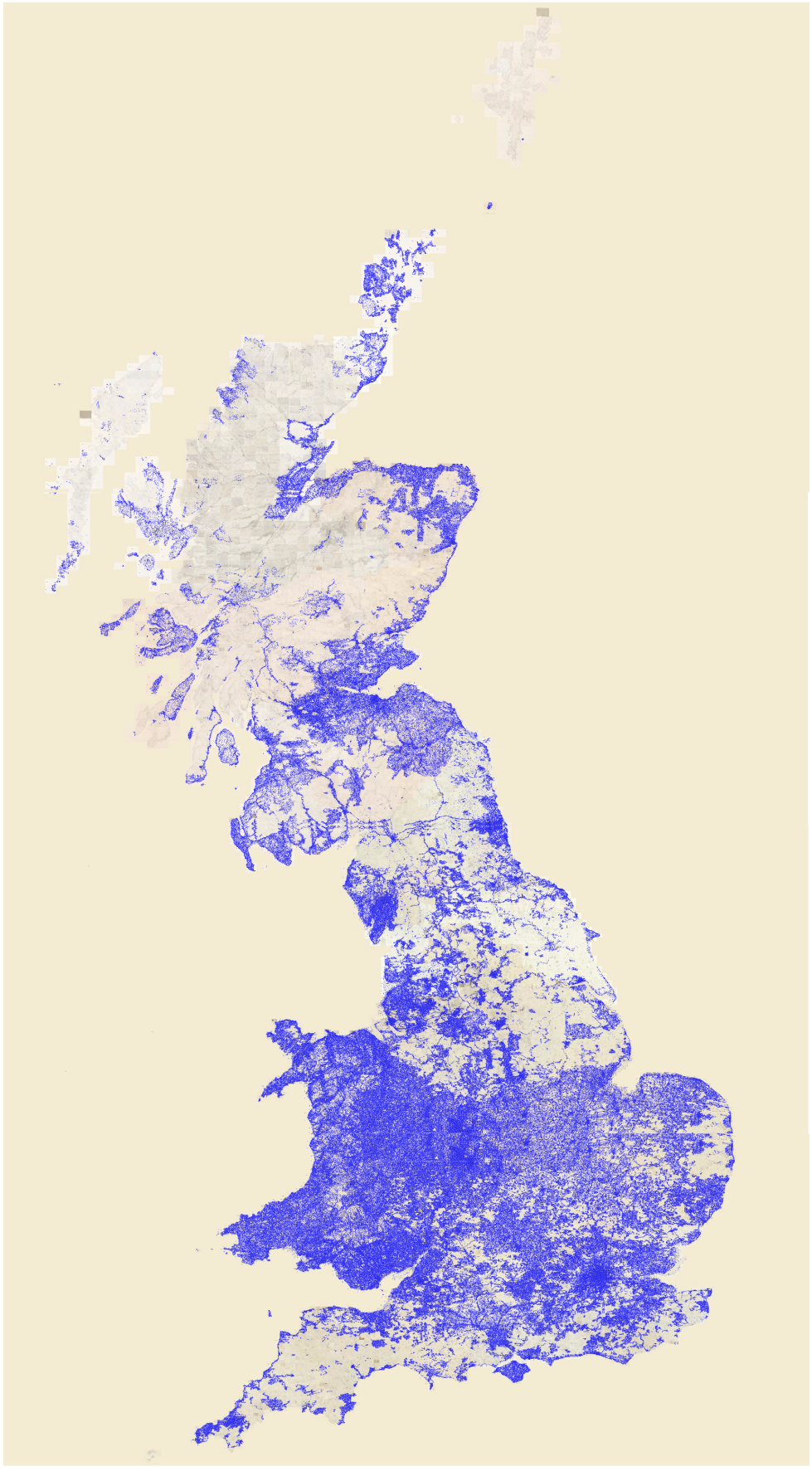

Figure 5. Maps showing progress of transcription: (c) 24th January 2017.

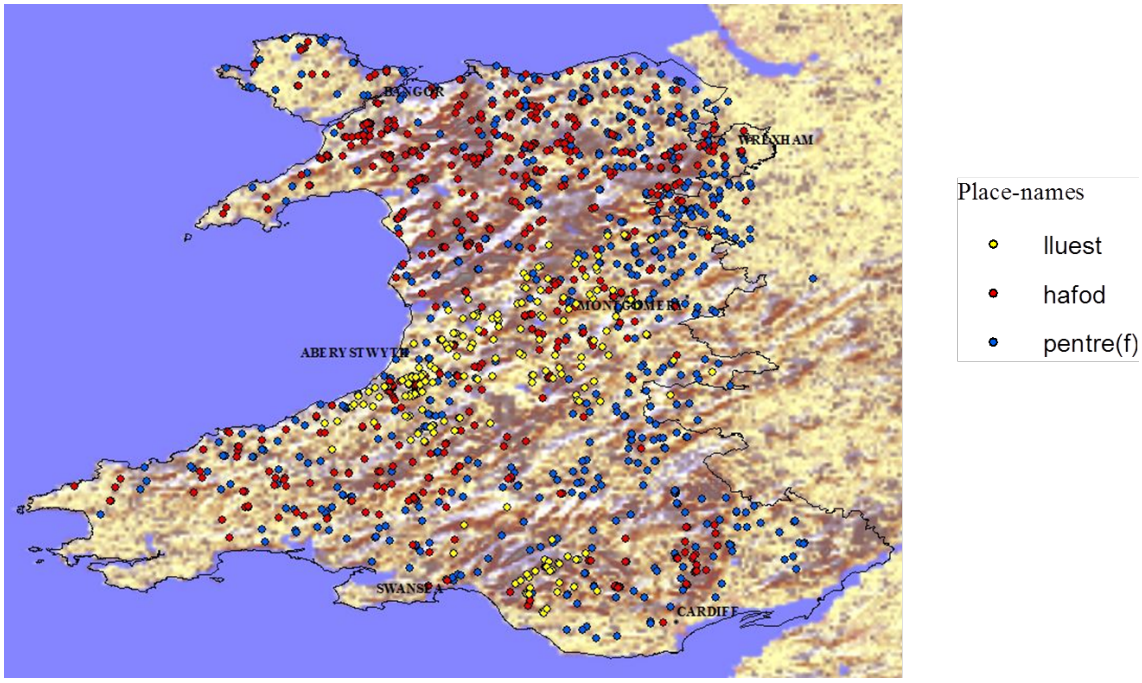

Figure 6. Distribution of *lluest*, *hafod* and *pentre* place-names in Wales